Phenotype ontologies for mouse and man: bridging the semantic gap

Schofield, Paul N; Gkoutos, Georgios V; Gruenberger, Michael; Sundberg, John P; Hancock, John M

Published in:
Disease Models & Mechanisms (DMM)

DOI:
10.1242/dmm.002790

Publication date:
2010

Citation for published version (APA):
Schofield, P. N., Gkoutos, G. V., Gruenberger, M., Sundberg, J. P., & Hancock, J. M. (2010). Phenotype ontologies for mouse and man: bridging the semantic gap. *Disease Models & Mechanisms (DMM)*, *3*(5-6), 281-289. https://doi.org/10.1242/dmm.002790

Disease Models & Mechanisms 3, 281-289 (2010) doi:10.1242/dmm.002790

Phenotype ontologies for mouse and man: bridging the semantic gap

Paul N. Schofield[1,2,*], Georgios V. Gkoutos[3], Michael Gruenberger[1], John P. Sundberg[2] and John M. Hancock[4]

A major challenge of the post-genomic era is coding phenotype data from humans and model organisms such as the mouse, to permit the meaningful translation of phenotype descriptions between species. This ability is essential if we are to facilitate phenotype-driven gene function discovery and empower comparative pathobiology. Here, we review the current state of the art for phenotype and disease description in mice and humans, and discuss ways in which the semantic gap between coding systems might be bridged to facilitate the discovery and exploitation of new mouse models of human diseases.

Mouse models of human diseases

The value of the mouse as a model for human disease has become firmly established as new mutants are repeatedly validated as models of human disease and, increasingly, the similarities in the pathobiology of the two species provide new insights into disease mechanisms and aetiologies (Peters et al., 2007; Rosenthal and Brown, 2007; Justice, 2008; Brown et al., 2009). Mutant strains derived from hypothesis-driven research are now being augmented by large-scale mutagenesis efforts that are being undertaken worldwide (Brown et al., 2009). Following the successful phenotype-driven *N*-ethyl-*N*-nitrosourea (ENU) mutagenesis projects, the products of which are still being analyzed, large-scale gene knockout programmes have been established to provide the mutant embryonic stem (ES) cells and mice that are needed to discover the functions of all of the protein-coding genes in the mouse genome. The International Knockout Mouse Consortium, (IKMC; www.knockoutmouse.org) (International Mouse Knockout Consortium et al., 2007), composed of four international partners (EUCOMM, KOMP, NorCOMM and TIGM), is currently producing large collections of targeted and gene-trapped mouse mutants. Currently 13,374 genes have been knocked out from a target number close to 24,000. More than 500 mouse lines are expected to be systematically phenotyped within the next five years using standardised phenotyping procedures developed by the EUMORPHIA (European Union Mouse Research for Public Health and Industrial Applications) and EUMODIC (The European Mouse Disease Clinics) consortia (www.eumodic.org) (Brown et al., 2005).

The mutagenesis efforts are not the only new sources of large amounts of systematic phenotyping data. The Shock-Ellison Medical Foundation-funded mouse aging programme at the Jackson Laboratory (http://agingmice.jax.org/index.html) is keeping mice from 31 different strains for the entirety of their natural life span to generate a huge volume of age-dependent phenotype data covering physiology, pathology and gene expression (Yuan et al., 2009). Longitudinal, cross-sectional and targeted studies of these mice provide interesting insights into the pathophysiology of aging. By using the high-resolution single nucleotide polymorphism (SNP) maps that are now available, these data will generate new gene/phenotype associations for many age-related conditions and complex traits.

To make the best use of the sheer volume and depth of the emerging mouse phenotype data we need to be able to relate it to human 'phenotype' or disease data in a way that is amenable to computation; it is this challenge that we discuss here.

What is a phenotype?

The concept of a *phenotype* is used in a variety of ways, not all of which are compatible with each other. Descriptions of clinical diseases (signs and symptoms), pathological lesions and entities; summative disease nomenclature (e.g. syndromes); the appearance or behaviour of mutants; genetically determined traits of strains; and, at the molecular level, transcriptome and gene expression patterns all represent examples of the common understanding of the concept of *phenotype*. When defined properly, the *phenome* itself is all of the genetically determined traits manifested under the prevailing environmental conditions and a *phenotype* is an observable property of the organism in the specified environment. Another useful concept is the *phenoset*, which represents a group of phenotypes in the same individual (e.g. behaviour, cancer, adiposity) that, together, characterise it.

Phenotypes versus traits

The term *phenotype* is often used as a synonym for a *trait*, especially in the description of human disease. This leads to considerable confusion. In the development of ontologies, the distinction between traits and phenotypes is essential for logical clarity and, in line with other developers (e.g. Hughes et al., 2008), we adopt the following definitions. A trait is a heritable, specifically measurable or identifiable feature of an organism, which can be followed through the genetic segregation of one or more phenotypes – such as short legs or dark hair. The traits here are 'leg length' and 'hair colour'. 'Short legs' and 'dark hair' are phenotypes, which are properties that can be measured or categorised under given environmental conditions.

[1]Department of Physiology, Development and Neuroscience, and [3]Department of Genetics, University of Cambridge, Cambridge, CB2 3EG, UK
[2]The Jackson Laboratory, Bar Harbor, ME 04609, USA
[4]Bioinformatics Group, MRC Harwell, Harwell, Oxfordshire, OX11 0RD, UK
*Author for correspondence (ps@mole.bio.cam.ac.uk)

The importance of phenotype data

For the mouse, useful associations can be made between genotypes and phenotypes where the mutation is known; either from identification of the ENU-induced or insertional alleles, or from targeted mutations. Additionally, where haplotype analysis is possible between inbred strains, making the association between phenotype and genotype permits the association of phenotype differences with specific haplotypes or SNPs, which is invaluable for complex trait analysis. This is now facilitated greatly by SNP discovery using high-throughput sequencing (Nikolaev et al., 2009).

Phenotypes that are shared between humans and mice can help identify candidate genes for human diseases. For example, candidate disease genes within association intervals in human genetic mapping studies, e.g. genome-wide association studies, may be triaged by looking at phenotypes of genes within the orthologous interval in the mouse. Evolutionary conservation of gene co-expression patterns for closely related phenotypes allows candidate gene prioritisation and, apart from identifying mouse mutants that can act as models for human diseases, we now see instances where high-resolution phenotyping of the mouse generates novel insights into human conditions (Ishimori et al., 2006; Ackert-Bicknell et al., 2008; Lisse et al., 2008).

The development of a common framework to describe human diseases and similar phenotypes in model organisms is needed to integrate the huge amount of phenotypic and genetic data that is generated from clinical genetic studies and the analysis of mutant animals. The problem is how to construct such a harmonised framework starting from the existing, well-established, but fundamentally different, approaches to describing phenotypes in humans and mice.

Coding of phenotype data

Both mice and humans have been 'phenotyped' for many years. Phenotypic variation in mice was recognised by the ancient Chinese (Keeler and Fuji, 1937) more than 2000 years ago. The *Eh Yah* dictionary (1100 B.C.) has a special term for a 'mouse with the hair pattern of a leopard', which is maybe the first description of a spontaneous mutation in the endothelin type B receptor gene, such as piebald ($Ednrb^{s\text{-}l}$) (Lane, 1966), which shows a characteristic black spotting. The 'waltzing' phenotype, which is probably the result of vestibular defects that are similar to the familiar spontaneous *waltzer* mutations, e.g. $Cdh23^{v\text{-}5J}$, was valued by the Japanese. A treatise on 'The Breeding of Curious Varieties of Mice', was published in 1787 by Chobei Zeniya of Kyoto, Japan. In this work, the author describes the crossing of various types of fancy mice and identifiably mentions the albino, non-agouti, recessive piebald, lilac with pink-eye and other heritable phenotypes. Our interest in mouse phenotypes and their genetics has a rich history.

A medical classification of human disease, known as nosology, has been attempted many times; in antiquity by Hippocrates and Isidore of Seville, and then later by Carl Linnaeus, who undertook one of the first attempts at a modern systematic classification of disease on the basis of symptoms (von Linne and Schroeder, 1763). Although subsequent classification systems have successively replaced these, the current system of International Classification of Diseases (ICD) for humans (World Health Organisation, 2008), still works in a paradigm that Linnaeus would recognise. To date, however, there have been few systematic attempts to harmonise the description of abnormality or disease between different species.

Phenotypes are generally described in natural language, frequently using a mixture of unstructured terminologies and free text, with variations that are widely understood within specific disciplines. Qualitative data is represented using disparate data models and indexed with simple text descriptions. At worst, the descriptions used for human phenotypes reflect local informal term usage or domain-specific controlled vocabularies. At best, they use terms from internationally accepted frameworks such as the Unified Medical Language System (UMLS), Medical Subject Headings (MeSH), International Classification of Disease (ICD-9/10) or Systematised Nomenclature of Medicine Clinical Terms (SNOMED-CT) terminology. SNOMED and ICD-9 are designed and structured for use in a clinical context, and both UMLS and MeSH are predominantly designed for describing human diseases and therapies. The structure of these nomenclatures precludes their use for logical inference and, in many cases, the terms are etiologically or anatomically predicated in a way that cannot be used to describe disease in non-human organisms. The result is that the coding of human data using these large and complex terminologies is logically and semantically incompatible with the type of coding and nomenclature used for model organisms such as mice.

Since natural language is highly expressive, the range of information it can capture in phenotype descriptions is usually both deep and broad. For example the 'hoarse cry' in Opitz GBBB syndrome (OMIM: 145410), and the 'striking upslanting of the palpebral fissures, small nose with broad root, abnormally modelled ears, short neck with loose skin' in Opitz C syndrome (OMIM: 211750) are difficult to express as concisely in any other way. Thus, natural language is the most obvious medium in which to record and express phenotypes. However, it is hard to carry out computation on descriptions based on natural language, and the task suffers from the now often-rehearsed problems of ambiguity, semantic complexity and lack of structure. For example the term *hedgehog* can refer to one of several human or mouse genes; human or mouse gene products; a small mammal of the family Erinaceomorpha; or an arrangement of pineapple and cheese impaled on cocktail sticks. Disambiguation and semantic standardisation are vital but difficult to achieve.

The key to providing terminological clarity is to use far more formalised language sets than are provided by natural language. The bioinformatics community realised this more than a decade ago and has produced complex term hierarchies describing various areas of knowledge (gene properties, anatomies, etc.) where the terms are linked by relationships (e.g. part of, is a, derived from, etc.). These ontologies have provided computational tools to capture knowledge within a domain and to express it within a relational framework that can be used by a broad range of clinicians and scientists (see Box 1) (Bard and Rhee, 2004). The most important ontologies for describing human abnormalities are the Human Phenotype Ontology (HPO) (Robinson et al., 2008) and the Disease Ontology (DO) (Du et al., 2009; Osborne et al., 2009), whereas those for the mouse are the Mammalian Phenotype Ontology (MP) (Smith et al., 2005) and the Mouse Pathology Ontology (MPATH; www.obofoundry.org/cgi-bin/detail.cgi?id=mouse_pathology). All are members of the Open Biological Ontology (OBO) family

Box 1. Ontologies

An *ontology* is a formal conceptual representation of a domain of knowledge with the primary aim of creating a shared understanding of a domain and the relationships within that domain. It contains common defined symbols for the *concepts* within a domain and meaningful relationships between those concepts. These relationships permit inference – the propagation of meaning across the ontology.

Most biomedical ontologies are structured as simple hierarchies of information using *is_a* or *part_of* relationships. For example the big toe is a *part_of* the foot and the heart *is_a* thoracic organ. These hierarchies are termed *directed acyclic graphs* as cyclic relationships are not permitted, i.e. one term is not permitted to be the parent and child of another term, and the flow of meaning through the hierarchy is from the most-specific term to the least specific.

(Smith et al., 2007) and can be downloaded from the OBO foundry site (www.obofoundry.org/).

Mammalian phenotype ontology and the mouse pathology ontology

The Mouse Genome Informatics (MGI) databases (www.informatics.jax.org/) (Eppig et al., 2007; Bult et al., 2008) hold qualitative (categorical) data coded by the MP (Smith et al., 2005). The MP consists largely of 'pre-coordinated' terms (see Box 2) – i.e. terms that include, for example, severity qualifiers or anatomical locations – and currently contains 9861 concepts.

The MP is currently the most successful and readily applicable approach to describing a wide range of aspects of phenotype and disease using a set of carefully defined descriptive terms. The terminology effectively captures various abnormal phenotypes and processes, as well as summative diagnoses and other descriptors of phenodeviance, which is the deviance of a phenotype in an animal, or cohort of genetically identical animals, away from what is typical in a reference population. Phenodeviance includes abnormal values for characteristics such as weight, coat colour or blood metabolites. The upper level terms of the MP ontology include physiological systems, behaviour, developmental phenotypes and ageing, and below this level, physiological systems are divided into morphological and physiological phenotypes. Many disease manifestations can be coded readily by MP and currently there are 88,600 annotations of approximately 21,000 genotypes in the MGI database. MP is a classically structured, hierarchy-based ontology and is designed to enable phenotype databases to be searched in order to find mutations and alleles with specific phenotypes; allow gene clustering based on mutant phenotypes; and discover genes in related pathways or potential mouse models of human diseases.

MPATH was originally designed as a description ontology for images of mouse histopathology and is segmented into aspects of pathology that would be familiar to traditionally trained pathologists. The most recent release is fully defined and contains terms covering all of the major classes of pathological lesions (594 to date), with specific reference to the mouse. These classes are arranged as a hierarchy within a directed acyclic graph (DAG), six levels deep, using the *is_a* relationship (e.g. a Harderian gland carcinoma *is_a* glandular tumour, *is_a* neoplasm) with each item having an MPATH ID that can be used for database interoperability and analysis. Many tissue responses are common to multiple anatomical sites and, as far as possible, the redundancy of specifying a particular response in multiple tissues has been avoided. The additional topographical or anatomical information for each image comes from the curatorial creation of cross-products with an appropriate anatomy ontology such as MA, the mouse adult anatomy (Hayamizu et al., 2005). For example, colon adenocarcinoma=[MPATH; 0000268 (Adenocarcinoma) + MA; 0000335 (Colon)]. The use of cross-products prevents the combinatorial explosion that causes 'ontology bloat' in poorly structured ontologies – the inclusion in the ontology of all possible pre-composed variations of instances of an entity (see Box 2).

Box 2. Pre-composition and post-composition

Also known as pre-coordination methodology, pre-composition uses a predefined set of phenotype terms created in advance by the ontology developer and combines, for example, the entity, say 'big toe', and the quality of that entity, say '[large big toe]'.

Also known as post-coordination methodology, post-composition involves construction of phenotype description at the time of annotation. In this case, there would be a term for 'big toe' and a term for 'large', and the post-composed term would combine these: ['big toe' + 'large']. This avoids, for example, the combinatorial explosion that is evident when the big toes might have many attributes that could also describe other toes, e.g. ['small big toe'], ['blue big toe'], ['short big toe'], ['short little toe'], ['large little toe'], etc.

Human disease ontology and human phenotype ontology

The full DO and its cut-down version, DO-lite (Du et al., 2009; Osborne et al., 2009), are based on ICD-9 and referenced to UMLS and SNOMED-CT. The full version contains 11,961 terms in the form of a hierarchy, of which 4399 terms are internal nodes lying up to 16 levels deep. HPO, the human phenotype ontology, was however derived from the terms found in the 'clinical synopsis' section of Online Mendelian Inheritance in Man (OMIM; www.ncbi.nlm.nih.gov/omim/) (Hamosh et al., 2005), and therefore covers largely monogenic diseases with mendelian inheritance. Although a hugely valuable resource, OMIM is not structured formally and the terminology used does not follow any consistent pattern. The construction of the HPO therefore represents a major improvement in the utility of OMIM and provides immediate structured genotype annotation to all of the 4779 annotated diseases. Both GeneRIFs (Mitchell et al., 2003) and GeneReviews (www.ncbi.nlm.nih.gov/projects/GeneTests/static/about/content/reviews.shtml) are additional useful sources of genotype/phenotype data but again are textual resources only.

The use of ontologies for recording human phenotypes is in its infancy and it is fair to say that the mouse research community has been much more pro-active in accepting and implementing standard terminologies than that of the human. The call for a human phenome project in 2003 (Freimer and Sabatti, 2003) with emphasis on the need for standards and international integration has not yet met with a concerted response, and with regard to human phenotypes and traits, there is an uncoordinated scatter of human phenotype and trait data throughout databases and resources across the world. Much human phenotype data relates to disease and its predisposition, and is largely captured with free text. In the best situations, it is coded using clinical

informatics formalisms such as ICD-9/10 or SNOMED-CT. These systems are structured, unambiguous and widely accepted but suffer from being highly pre-composed (e.g. aetiologically and anatomically predicated), and are not organised in such way as to support inference or computer reasoning. Nevertheless, one great advantage is that tools and resources such as UMLS and MetaMap (Bodenreider, 2004) are available for using ICD and SNOMED coding systems. These provide synonyms, cross-references and mark-up facilities, which are of assistance in comparing data between databases and within literature records, and have recently been used in crossing species boundaries (see below) (Marquet et al., 2007).

Human genetic databases may be divided into core databases and locus-specific databases (LSDB). Core databases attempt to provide data on all pathological variation and its consequences, for example, the human gene mutation database (HGMD) (Stenson et al., 2008), which uses a local controlled vocabulary. LSDBs, by contrast, focus on one gene or locus respectively [for discussion, see Patrinos and Brookes (Patrinos and Brookes, 2005)]. The genetic association database (GAD) (Becker et al., 2004) contains associations between complex diseases and disorders and individual human genes curated from the literature; here, diseases are categorised using a controlled vocabulary drawn from MeSH terms. Quantitative data sets on human populations are held by the database of genotypes and phenotypes, DBGaP (Mailman et al., 2007), and again are indexed in a largely unstructured way through MeSH-defined terms. The human genome variation database, HGVbase G2P (Thorisson et al., 2009), is one of the most useful collections of genotype/phenotype associations, although it uses only a local controlled vocabulary to record phenotype data.

The consequence of the terminology 'Babel' in human clinical databases is that text mining is often the only approach to extract information from these resources (Perez-Iratxeta et al., 2002; Hristovski et al., 2005; van Driel et al., 2006). Text mining is fraught with problems, including issues of semantics, over-representation of common phenotypes and insufficient granularity.

Misinterpretation of the literature, combined with inaccurate database curation, can generate misleading hypotheses through implied disease orthology. However, the following example of the mouse hairless gene and its incorrect link to the complex polygenic disease known as alopecia universalis in humans shows that more considered analysis of such errors can ultimately create a much greater understanding of a particular disease. The hairless phenotype and its more severe form, known as rhino (short for rhinoceros), were first described in mice in 1856 (Gaskoin, 1856). The human homologue, atrichia with papules, or as it later became known as, papular atrichia, was first described in 1954, nearly 100 years later (Damste and Prakken, 1954). The link between the mouse and human disease was made some 30 years afterwards (Sundberg et al., 1989; Sundberg, 1994). The hairless gene was traditionally linked to a simple, recessively inherited form of alopecia universalis based on a curation call in the OMIM entry (OMIM: 203655) (Ahmad et al., 1998). The OMIM designation was based on morphologic diagnosis; a total lack of hair in patients with an autosomal recessive pattern of inheritance. Alopecia universalis is actually a well-characterised, complex genetic-based autoimmune skin disease in both humans (Martinez-Mir et al., 2007) and mice (Sundberg et al., 2004). Although this mismatch was initially of great concern (Sundberg et al., 1999), it subsequently led to a much better understanding of papular atrichia. Many mutations have now been identified in the human hairless gene, as well as in rodents and non-human primates (Panteleyev et al., 1998; Ahmad et al., 2002).

Crossing the species divide; granularity and specificity

Accurate phenotype descriptions can discover new relationships between genes and phenotypes, and new functions for previously uncharacterised genes and alleles. A good example is PhenomicDB (Groth et al., 2007), which contains one of the most wide-ranging cross-species datasets on gene/phenotype associations. This database combines data from OMIM, the Mouse Genome Database (MGD), WormBase, FlyBase, the Comprehensive Yeast Genome Database (CYGD), the Zebrafish Information Network (ZFIN), and the MIPS *Arabidopsis thaliana* database (MAtDB). Groth et al. (Groth et al., 2008) queried the resulting PhenomicDB 'warehouse' that was created by using a text-mining approach, and which generated a summary phenotypic statement for each gene, then clustered the statements to produce what Oti and Brunner (Oti and Brunner, 2007) have termed 'Phenoclusters' – a group of genes with overlapping phenotypes, which may then be used for the discovery of new disease or functional associations. This phenotype-driven approach to the discovery of gene function has distinct advantages over the gene-driven approach to phenotype prediction because, although many closely related phenotypes are caused by mutations in different genes whose gene products interact directly or are on the same pathway, mutations in the same gene can have diverse phenotypic outcomes depending on which function of a multifunctional gene product is compromised. Several related disease candidate gene discovery approaches have been developed (for examples, see Tiffin et al., 2006; van Driel and Brunner, 2006; Oti and Brunner, 2007). However, in the absence of systematic coding, all of these approaches depend to a greater or lesser extent on text mining from their data sources, and making use, at best, of UMLS and MeSH terms in abstracts and database phenotype fields. Despite impressive results from many of these approaches, it is clear that a standardised description of phenotypes and diseases would greatly increase the power and specificity of cross-species data mining.

A key problem is the assumption that the currently dominant paradigm for disease conceptualisation, based on clinical medicine, is useful for biomedical science applications. It is a mistake to assume that the human 'phenome' is a list of 'diseases' that form more or less distinct entities. The realisation that diseases of separate genetic aetiology may share similar phenotypes may seem obvious, but it is only recently that this has generated attention. Work by Brunner and others (Oti and Brunner, 2007; Oti et al., 2008) demonstrated that shared aspects of phenotype may be viewed as a proxy for a common underlying pathogenetic mechanism, and that this mechanism may be shared by dysfunction of a group of genes whose products either interact or are on the same functional pathway. This 'modularity' of phenotypes should not come as a surprise, but it makes the formulation of a new

concept of disease description all the more urgent. The generation of phenoclusters depends on the ability to code phenotypes in as granular a way as possible. This approach was used originally in making gene/phenotype associations in RNA interference (RNAi)-generated phenotypes (by our definition, *phenosets*) in *C. elegans*, where each was expressed as a combination of 45 phenotypic features, enabling clustering of functionally related genes (Piano et al., 2002).

The use of a phenotype-driven approach to discover new information about gene/phenotype relationships *within* a species requires a sufficiently high level of specificity and granularity to discriminate between closely related phenotypes with overlapping components. This is particularly true of complex traits. Joy and Hegele (Joy and Hegele, 2008) provide an excellent discussion of the problems caused by the inaccuracy and variability of definitions in the context of metabolic syndrome and the resulting problems with candidate gene association and linkage studies. Description problems inhibited gene association studies in X-linked mental retardation, where there are insufficient phenotypic features to 'unbundle' non-syndromic cases in gene association studies (Ropers and Hamel, 2005).

The requirement for 'deep phenotyping' using well-defined criteria is clearly important in human gene association studies. It is also crucial if human phenotypes are to be compared with those from model organisms. The deficiency in cross-species interoperability of phenotype description formalisms is well demonstrated by the analysis of cross-species phenoclustering that was carried out using PhenomicDB by Groth et al. (Groth et al., 2007), discussed above. More than 90% of the clusters they generated contained genes from a single species and there was a tendency for genes to fall into species-specific clusters. They interpret this as an indication that the terminology used to describe phenotype in each species fails to cross the species barrier, even though many phenotypes clearly have their equivalents between species. It is therefore clear that, if our aim is to understand the underlying processes and genetic aetiology through using model organisms, we need a change in the way in which diseases and phenotypes are described.

Bridging ontologies

The ontologies described earlier were all developed for particular species and, like many other controlled vocabularies, are not readily interoperable for cross-species queries, for example, between different genotype/phenotype databases. Semantic inconsistency and anatomical incompatibility, together with different traditions of disease description in different organisms, prevent the matching of phenotype ontologies either lexically or conceptually.

Two related problems impede the bridging of different ontologies that have been derived for either the same or separate species. None of the ontologies or controlled vocabularies for describing disease is truly orthogonal (generally used in this context to mean complementary and non-redundant), although they were designed to cover the same area of knowledge, for example DO and HPO. This means that, even within a species, the terminology used and the underlying structure may be different. For example, the term 'melanoma in situ' is used within SNOMED-CT and MPATH to represent a potentially cancerous lesion, whereas the National Toxicology Program (NTP) Toxicology Data Management System (TDMS) pathology code table for microscopic lesions (http://hazel.niehs.nih.gov/user_spt/pct_terms.htm) defines only 'melanoma benign' and 'melanoma malignant'. Similarly, only the NTP TDMS pathology code table and the SNOMED-CT vocabularies define a benign melanoma term ('melanoma benign' and 'benign melanocytic neoplasm', respectively). HPO provides 'especially prone to malignant melanoma', 'malignant intraocular melanoma' and 'malignant melanoma'. HPO does not address pre-neoplastic or benign lesions, but provides an anatomically predicated version and a predisposition syndrome. DO provides 96 melanoma terms, many of which are pre-composed and are both anatomically predicated and include morphological and prognostic qualifiers. Interestingly, there is no term for the pre-neoplastic lesion. MP only contains the anatomically predicated 'intraocular melanoma'. Even this superficial comparison shows that comparing the data coded to each of these ontologies is very difficult and impossible to do automatically using simple lexical matching. A major problem is the use of complex pre-composed terms (see Box 2). In comparison to this, the issue of species-specific lesions, for example, as is found when comparing mouse haematopoietic neoplasms with those in humans, is relatively easy to deal with (Kogan et al., 2002; Morse et al., 2002). Making use of the subsumption (incorporation of a term into a higher order or parental category) that is available within an ontology permits relation of species-specific variants through a common parent. For example, the mouse small T-cell lymphoma (STL), which probably has no counterpart in the human (Morse et al., 2002), can be classified as a 'mature T-cell neoplasm' – a parent category that is common to human and mouse malignancies. Searching a database of mouse and human tumours using an ontology, where STL *is_a* 'mature T-cell neoplasm', would recover any human data coded to the 16 'mature T-cell neoplasms' listed in ICD.

One approach to bridging the nomenclature gap between species is to make use of the UMLS resource of the National Libraries of Medicine (NLM). The UMLS thesaurus (Bodenreider, 2004) is a large and well-curated resource of terms and synonyms that can be used for semantic mapping between terminologies. This was used by Osborne et al. (Osborne et al., 2009) to annotate the human genome to the DO ontology and has proved a valuable approach to cross-mapping the DO, MP and MPATH (Marquet et al., 2007). However, apart from what might be described as the 'straightforward' compatibility problem, there is a more complex problem that needs to be considered: that of the composition of disease terms themselves.

An alternative way to represent phenotype: the E+Q approach

It is clear from the preceding discussion that a major problem in describing phenotypes and diseases is that many of the terms that are commonly used to describe them are complex and subsume a multitude of meanings. This is both a problem for cross-linking phenotype and disease, and restrictive computationally. The MP ontology, for example, only allows the description of abnormal phenotypes and does not allow quantitative descriptions. An alternative approach is to break down complex pre-composed terms into their constituent logical parts, an approach known as the E+Q (entity plus quality) approach, which is used in the capture of raw mouse phenotype data

(Bard and Rhee, 2004; Gkoutos et al., 2004; Gkoutos et al., 2005; Mungall et al., 2007; Beck et al., 2009). The E+Q syntax uses a combination of relevant descriptive ontologies. It represents entities (E), such as anatomical structures or chemical compounds, using ontologies such as MA, the Foundational Model of Anatomy (FMA) (Rosse and Mejino, 2003) and Chemicals of Biological Interest (CheBI) (Degtyarenko et al., 2008), etc., and represents the qualities (Q) inhering in the entities, such as colour, size or shape using the Phenotype and Trait Ontology, PATO (www.obofoundry.org/cgi-bin/detail.cgi?id=quality). The combination of E and Q terms can then be used to represent both traits (e.g. 'tail+length') or phenotypes (e.g. 'tail+long'); within PATO, 'long' in this example is a child of 'length', so that the trait is implicit in the phenotype (Gkoutos et al., 2004; Gkoutos et al., 2005; Beck et al., 2009). The basic E+Q syntax can be extended to increase expressivity to include E_2, which is an additional optional entity type for relational qualities, and the modifier M: E+Q+E_2+M.

The E+Q approach is referred to as 'post-composition', reflecting the composition of compound terms from components. It is used in the EuroPhenome mouse phenotype database (www.europhenome.org/) to describe raw phenotype data from high-throughput phenotyping experiments (Beck et al., 2009; Morgan et al., 2010), and in some model organism databases such as ZFIN and FlyBase (Drysdale, 2008; Sprague et al., 2008). For example, to describe the phenotype of *Sox9* mutants, MGI uses the pre-composed term MP:0005587 (abnormal Meckel's cartilage) and ZFIN uses the E+Q approach – entity, ZFA:0001205 (Meckel's cartilage); quality, PATO:000587 (decreased size).

The E+Q approach can be used to provide a 'logical definition' of a pre-composed ontology term. Applying a decomposition process to pre-composed terms in principle allows terms with different names to be linked via shared logical definitions, a process that could be used to link phenotypes across species or, in principle, phenotypes to diseases. Using this approach, Mungall and co-workers (Mungall et al., 2010) recently reported the association of 8285 classes from four species-specific ontologies to E+Q definitions, using a cross-species upper level ontology of anatomy, Uberon (Haendel et al., 2009). Leveraging the E+Q definitions that were available for mouse, human and zebrafish phenotypes, Washington et al. (Washington et al., 2009) have been able to identify orthologous and biologically relevant genes on the basis of E+Q phenotype similarity, matching within and between species for a defined test set of genes, thereby validating the approach.

The relationship between pre- and post-composed ontologies is additionally advantageous as pre-composed ontologies, such as MP, are 'human-readable', whereas post-composed ontologies are better for computational analysis. An example of this is the EuroPhenome database (Beck et al., 2009; Morgan et al., 2010). Here, quantitative parameters for specific phenotypic assays are stored in the database. Mutant cohorts are then compared with control cohort data and statistically abnormal lines are annotated dynamically to E+Q statements of phenodeviance using preset parameters. Logical definitions then allow E+Q statements to be translated into pre-composed MP terms. Both quantitative and qualitative data can be represented in this way, and representation in MP allows the data to be queried in a consistent and transparent way that offers a powerful paradigm for the annotation and computational analysis of mutant phenotype data.

Disaggregation of disease entities

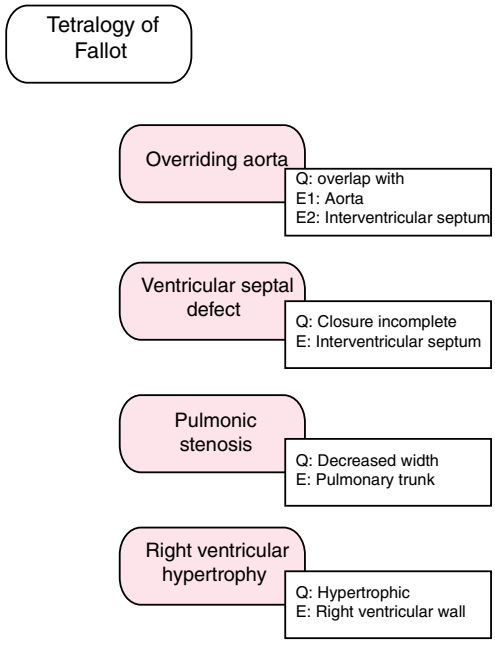
In principle, the ontology decomposition approach described above might be used to map phenotypes to diseases, and phenotypes between species. However, as discussed above, the term 'phenotype' is used to encompass a multitude of logically disparate entities. This is especially true with the terms that are commonly used to describe diseases in humans. Human diseases are complex collections of phenotypic observations and pathological processes, and a diagnosis involves establishing the presence of a set of phenotypes, which is often probabilistic. For example, Beckwith-Wiedemann syndrome is defined by the simultaneous presence in the proband of all of the three most common phenotypes (macroglossia, anterior abdominal wall defect and overgrowth), or two of these phenotypes combined with five of more than a dozen other manifestations (Elliott et al., 1994; Cooper et al., 2005).

As long as we do not have a formalism to capture the probabilistic phenotypic elements of diseases – often the underlying observations used by clinicians as diagnostic criteria – high-level disease terms will be difficult to use for detecting overlaps between diseases and between phenotypes in different species. Additional elements also need to be captured to accurately record aspects of genetic disease that are used for differential diagnosis and stratification, such as the mode of inheritance, penetrance, pleiotropy, expressivity and progression. This is a challenge for the E+Q framework.

A first step towards a solution is to disaggregate disease terms into individual phenotypic components, which, in combination, make up the disease entity. An example of this is shown in Fig. 1, using the HPO. Here, the congenital heart defect tetralogy of Fallot, which is very difficult to render into a satisfactory E+Q statement directly, is broken down into its constituent endophenotypes, which are then amenable to E+Q definition. With the provision of a bridging anatomy ontology, the remaining terms used in the E+Q statements are from PATO and are species agnostic, allowing species-specific phenotype data to be traversed readily.

As an illustration of the potential utility of this disaggregation approach, we set out to search MGI for models of the tetralogy of Fallot. MP does not contain the term 'tetralogy of Fallot', but searching MGI with the intersection of MP:0000273 (overriding aorta), MP:0000486 (abnormal pulmonary trunk morphology), MP:0000276 (heart right ventricle hypertrophy) and MP:0008823 (abnormal membranous ventricular septum morphology), yields the homozygous knockout of hairy/enhancer-of-split related with YRPW motif 2 (*Hey2*^{tm1Uts}), which is already annotated as a model for the tetralogy of Fallot in OMIM, and homozygous knockout of polyhomeotic-like 1 (*Phc1*^{tm1Os}), which has not previously been linked.

This is a relatively straightforward example. However, disease description is a complex domain and disaggregating disease terms requires expert input if the disaggregations are to accurately reflect the clinical nature of the disease. Although automatic approaches, such as the ones used in the HPO, are a great advance, the cooperation of experts in individual disease areas is needed to produce a well-founded, 'post-composed' disease ontology.



[Term]id: HP:0001636 ! Tetralogy of Fallot
intersection_of: PATO:0000001 ! Quality
intersection_of: has_part HP:0002623 ! Overriding aorta
intersection_of: has_part HP:0001629 ! Ventricular septal defect
intersection_of: has_part HP:0001642 ! Pulmonic stenosis
intersection_of: has_part HP:0001667 ! Right ventricular hypertrophy

[Term]id: HP:0002623 ! Overriding aorta
intersection_of: PATO:0001590 ! overlap with
intersection_of: inheres_in FMA:3734 ! Aorta
intersection_of: towards FMA:7135 ! Membranous interventricular septum

[Term]id: HP:0001629 ! Ventricular septal defect
intersection_of: PATO:0000609 ! closure incomplete
intersection_of: inheres_in FMA:7133 ! Interventricular septum

[Term]id: HP:0001642 ! Pulmonic stenosis
intersection_of: PATO:0000599 ! decreased width
intersection_of: inheres_in FMA:8615 ! Pulmonary arterial trunk

[Term]id: HP:0001667! Right ventricular hypertrophy
intersection_of: PATO:0000584! hypertrophic
intersection_of: inheres_in FMA:9533! Right ventricular wall

Fig. 1. Endophenotype disaggregation of the tetralogy of Fallot (OMIM:187500). The syndrome is broken down into its component endophenotypes, each of which appears independently both in a clinical context and in HPO. Each endophenotype is then logically defined using anatomy terms from the FMA qualified with PATO. The description of the resulting set of entities and definitions is given in OBO syntax (www.geneontology.org/GO.format.obo-1_2.shtml) on the right. The tetralogy can be found in isolation, both sporadically and in the context of a heritable predisposition, or as a component of another syndrome, in which case the definitions here would become nested in that of the higher level syndrome, such as DiGeorge syndrome (OMIM:188400) or velocardiofacial syndrome (OMIM:192430).

Conclusions

Ontologies and description frameworks for capturing data on disease and phenotype are essential tools to support mouse functional genomics, and in a broader context, for the assignation of functions to genes. The tools that are currently available are still in the early stages of development and may need to be applied in new ways to fully serve the requirements of cross-species phenotype mapping. Even a preliminary attempt to implement existing ontologies in the E+Q framework demonstrates the need for more terms to describe measured entities, both in humans and in mice, and for example a mammalian trait ontology would be of great utility. Another area in need of development is that of the non-anatomical phenotype traits, notably behaviour. With respect to the human, it will not always be possible to obtain or record measurements with the same completeness or precision as with mice in a laboratory setting, although some clinical biobanking projects approach this, and in many cases phenotype description from the literature will inevitably be only qualitative, if only because it constitutes legacy data. The power of the decompositional approach is that it is applicable to both qualitative and quantitative data and, in either, lends itself to computational analysis. The difficulty and amount of labour necessary to implement effective cross-species ontologies is daunting, but success will yield valuable insights from model organisms.

ACKNOWLEDGEMENTS

This work was funded by the Commission of the European Community Contract number LSHG-CT-2006-037811; CASIMIR. J.P.S. acknowledges support of the National Institutes of Health (CA089713) and the Ellison Medical Foundation. The authors thank Prof. Jonathan Bard, Prof. Janan Eppig, Dr Peter Robinson and Dr Anita Burgun for discussions and for helpful comments on the manuscript. Deposited in PMC for release after 12 months.

COMPETING INTERESTS

The authors declare no competing financial interests.

REFERENCES

Ackert-Bicknell, C. L., Demissie, S., Marin de Evsikova, C., Hsu, Y. H., DeMambro, V. E., Karasik, D., Cupples, L. A., Ordovas, J. M., Tucker, K. L., Cho, K. et al. (2008). PPARG by dietary fat interaction influences bone mass in mice and humans. *J. Bone. Miner. Res.* **23**, 1398-1408.

Ahmad, W., Faiyaz, U., Brancolini, V., Tsou, H. C., Haque, S. u., Lam, H., Alta, V. M., Owen, J., deBlaquiere, M., Frank, J. et al. (1998). Alopecia universalis associated with a mutation in the human hairless gene. *Science* **279**, 720-724.

Ahmad, W., Ratteree, M. S., Panteleyev, A. A., Aita, V. M., Sundberg, J. P. and Christiano, A. M. (2002). Atrichia with papular lesions resulting from mutations in the rhesus macaque (Macaca mulatta) hairless gene. *Lab. Anim.* **36**, 61-67.

Bard, J. B. and Rhee, S. Y. (2004). Ontologies in biology: design, applications and future challenges. *Nat. Rev. Genet.* **5**, 213-222.

Beck, T., Morgan, H., Blake, A., Wells, S., Hancock, J. M. and Mallon, A. M. (2009). Practical application of ontologies to annotate and analyse large scale raw mouse phenotype data. *BMC Bioinformatics* 10 **Suppl. 5**, S2.

Becker, K. G., Barnes, K. C., Bright, T. J. and Wang, S. A. (2004). The genetic association database. *Nat. Genet.* **36**, 431-432.

Bodenreider, O. (2004). The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267-D270.

Brown, S. D., Chambon, P. and de Angelis, M. H. (2005). EMPReSS: standardized phenotype screens for functional annotation of the mouse genome. *Nat. Genet.* **37**, 1155.

Brown, S. D., Wurst, W., Kuhn, R. and Hancock, J. (2009). The functional annotation of mammalian genomes: the challenge of phenotyping. *Annu. Rev. Genet.* **43**, 305-333.

Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E. and Blake, J. A. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. *Nucleic Acids Res.* **36**, D724-D728.

Cooper, W. N., Luharia, A., Evans, G. A., Raza, H., Haire, A. C., Grundy, R., Bowdin, S. C., Riccio, A., Sebastio,

G., Bliek, J. et al. (2005). Molecular subtypes and phenotypic expression of Beckwith-Wiedemann syndrome. *Eur. J. Hum. Genet.* **13**, 1025-1032.
Damste, J. and Prakken, J. R. (1954). Atrichia with papular lesions: a variant of congenital ectodermal dysplasia. *Dermatologica* **108**, 114-121.
Degtyarenko, K., de Matos, P., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcantara, R., Darsow, M., Guedj, M. and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Res.* **36**, D344-D350.
Drysdale, R. (2008). FlyBase: a database for the Drosophila research community. *Methods Mol. Biol.* **420**, 45-59.
Du, P., Feng, G., Flatow, J., Song, J., Holko, M., Kibbe, W. A. and Lin, S. M. (2009). From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics* **25**, i63-i68.
Elliott, M., Bayly, R., Cole, T., Temple, I. K. and Maher, E. R. (1994). Clinical features and natural history of Beckwith-Wiedemann syndrome: presentation of 74 new cases. *Clin. Genet.* **46**, 168-174.
Eppig, J. T., Blake, J. A., Bult, C. J., Richardson, J. E., Kadin, J. A. and Ringwald, M. (2007). Mouse genome informatics (MGI) resources for pathology and toxicology. *Toxicol. Pathol.* **35**, 456-457.
Freimer, N. and Sabatti, C. (2003). The human phenome project. *Nat. Genet.* **34**, 15-21.
Gaskoin, J. S. (1856). On a peculiar variety of Mus musculus. *Proc. Zool. Soc. Lond.* **24**, 38-40.
Gkoutos, G. V., Green, E. C. J., Mallon, A.-M., Hancock, J. M. and Davidson, D. (2004). Building mouse phenotype ontologies. *Pac. Symp. Biocomputing* **9**, 178-189.
Gkoutos, G. V., Green, E. C. J., Mallon, A.-M., Hancock, J. M. and Davidson, D. (2005). Using ontologies to describe mouse phenotypes. *Genome Biol.* **6**, R8.
Groth, P., Pavlova, N., Kalev, I., Tonov, S., Georgiev, G., Pohlenz, H. D. and Weiss, B. (2007). PhenomicDB: a new cross-species genotype/phenotype resource. *Nucleic Acids Res.* **35**, D696-D699.
Groth, P., Weiss, B., Pohlenz, H. D. and Leser, U. (2008). Mining phenotypes for gene function prediction. *BMC Bioinformatics* **9**, 136.
Haendel, M., Gkoutos, G. V., Lewis, S. E. and Mungall, C. (2009). Uberon: towards a comprehensive multi-species anatomy ontology. In *International Consortium of Biomedical Ontology: 2009*. Buffalo, New York: Nature Proceedings.
Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A. and McKusick, V. A. (2005). Online mendelian inheritance in man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucleic Acids Res.* **33**, D514-D517.
Hayamizu, T. F., Mangan, M., Corradi, J. P., Kadin, J. A. and Ringwald, M. (2005). The adult mouse anatomical dictionary: a tool for annotating and integrating data. *Gen. Biol.* **6**, R29.
Hristovski, D., Peterlin, B., Mitchell, J. A. and Humphrey, S. M. (2005). Using literature-based discovery to identify disease candidate genes. *Int. J. Med. Inform.* **74**, 289-298.
Hughes, L. M., Bao, J., Hu, Z. L., Honavar, V. and Reecy, J. M. (2008). Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species. *J. Anim. Sci.* **86**, 1485-1491.
International Mouse Knockout Consortium, Collins, F. S., Rossant, J. and Wurst, W. (2007). A mouse for all reasons. *Cell* **128**, 9-13.
Ishimori, N., Li, R., Walsh, K. A., Korstanje, R., Rollins, J. A., Petkov, P., Pletcher, M. T., Wiltshire, T., Donahue, L. R., Rosen, C. J. et al. (2006). Quantitative trait loci that determine BMD in C57BL/6J and 129S1/SvImJ inbred mice. *J. Bone Miner. Res.* **21**, 105-112.
Joy, T. and Hegele, R. A. (2008). Genetics of metabolic syndrome: is there a role for phenomics? *Curr. Atheroscler. Rep.* **10**, 201-208.
Justice, M. J. (2008). Removing the cloak of invisibility: phenotyping the mouse. *Dis. Model. Mech.* **1**, 109-112.
Keeler, C. E. and Fuji, S. (1937). The antiquity of mouse variants in the Orient. *J. Hered.* **28**, 93-96.
Kogan, S. C., Ward, J. M., Anver, M. R., Berman, J. J., Brayton, C., Cardiff, R. D., Carter, J. S., de Coronado, S., Downing, J. R., Fredrickson, T. N. et al. (2002). Bethesda proposals for classification of nonlymphoid hematopoietic neoplasms in mice. *Blood* **100**, 238-245.
Lane, P. W. (1966). Association of megacolon with two recessive spotting genes in the mouse. *J. Hered.* **57**, 29-31.
Lisse, T. S., Thiele, F., Fuchs, H., Hans, W., Przemeck, G. K., Abe, K., Rathkolb, B., Quintanilla-Martinez, L., Hoelzlwimmer, G., Helfrich, M. et al. (2008). ER stress-mediated apoptosis in a new mouse model of osteogenesis imperfecta. *PLoS Genet.* **4**, e7.
Mailman, M. D., Feolo, M., Jin, Y., Kimura, M., Tryka, K., Bagoutdinov, R., Hao, L., Kiang, A., Paschall, J., Phan, L. et al. (2007). The NCBI dbGaP database of genotypes and phenotypes. *Nat. Genet.* **39**, 1181-1186.
Marquet, G., Mosser, J. and Burgun, A. (2007). A method exploiting syntactic patterns and the UMLS semantics for aligning biomedical ontologies: the case of OBO disease ontologies. *Int. J. Med. Inform.* **76 Suppl. 3**, S353-S361.
Martinez-Mir, A., Zlotogorski, A., Gordon, D., Petukhova, L., Mo, J., Gilliam, T. C., Londono, D., Haynes, C., Ott, J., Hordinsky, M. et al. (2007). Genomewide scan for linkage reveals evidence of several susceptibility loci for alopecia areata. *Am. J. Hum. Genet.* **80**, 316-328.
Mitchell, J. A., Aronson, A. R., Mork, J. G., Folk, L. C., Humphrey, S. M. and Ward, J. M. (2003). Gene indexing: characterization and analysis of NLM's GeneRIFs. *AMIA Annu. Symp. Proc.* **2003**, 460-464.
Morgan, H., Beck, T., Blake, A., Gates, H., Adams, N., Debouzy, G., Leblanc, S., Lengger, C., Maier, H., Melvin, D. et al. (2010). EuroPhenome: A repository for high-throughput mouse phenotyping data. *Nucleic Acids Res.* **38**, D557-D585.
Morse, H. C., 3rd, Anver, M. R., Fredrickson, T. N., Haines, D. C., Harris, A. W., Harris, N. L., Jaffe, E. S., Kogan, S. C., MacLennan, I. C., Pattengale, P. K. et al. (2002). Bethesda proposals for classification of lymphoid neoplasms in mice. *Blood* **100**, 246-258.
Mungall, C. J., Gkoutos, G. V., Smith, C. L., Haendel, M. A., Lewis, S. E. and Ashburner, M. (2010). Integrating phenotype ontologies across multiple species. *Genome Biol.* Jan 8 [Epub ahead of print] [doi:10.1186/gb-2010-11-1-r2].
Mungall, C. M., Gkoutos, G. V., Washington, S. E. and Lewis, S. E. (2007). Representing Phenotypes in OWL. In *Proceedings of the OWLED Workshop on OWL: Experience and Directions* (eds C. Golbreich, A. Kalyanpur and B. Parsia). Innsbruck, Austria.
Nikolaev, S. I., Iseli, C., Sharp, A. J., Robyr, D., Rougemont, J., Gehrig, C., Farinelli, L. and Antonarakis, S. E. (2009). Detection of genomic variation by selection of a 9 mb DNA region and high throughput sequencing. *PloS One* **4**, e6659.
Osborne, J. D., Flatow, J., Holko, M., Lin, S. M., Kibbe, W. A., Zhu, L. J., Danila, M. I., Feng, G. and Chisholm, R. L. (2009). Annotating the human genome with Disease Ontology. *BMC Genomics* **10 Suppl. 1**, S6.
Oti, M. and Brunner, H. G. (2007). The modular nature of genetic diseases. *Clin. Genet.* **71**, 1-11.
Oti, M., Huynen, M. A. and Brunner, H. G. (2008). Phenome connections. *Trends Genet.* **24**, 103-106.
Panteleyev, A. A., Paus, R., Ahmad, W., Sundberg, J. P. and Christiano, A. M. (1998). Molecular and functional aspects of the hairless (hr) gene in laboratory rodents and humans. *Exp. Dermatol.* **7**, 249-267.
Patrinos, G. P. and Brookes, A. J. (2005). DNA, diseases and databases: disastrously deficient. *Trends Genet.* **21**, 333-338.
Perez-Iratxeta, C., Bork, P. and Andrade, M. A. (2002). Association of genes to genetically inherited diseases using data mining. *Nat. Genet.* **31**, 316-319.
Peters, L. L., Robledo, R. F., Bult, C. J., Churchill, G. A., Paigen, B. J. and Svenson, K. L. (2007). The mouse as a model for human biology: a resource guide for complex trait analysis. *Nat. Rev. Genet.* **8**, 58-69.
Piano, F., Schetter, A. J., Morton, D. G., Gunsalus, K. C., Reinke, V., Kim, S. K. and Kemphues, K. J. (2002). Gene clustering based on RNAi phenotypes of ovary-enriched genes in C. elegans. *Curr. Biol.* **12**, 1959-1964.
Robinson, P. N., Kohler, S., Bauer, S., Seelow, D., Horn, D. and Mundlos, S. (2008). The human phenotype ontology: a tool for annotating and analyzing human hereditary disease. *Am. J. Hum. Genet.* **83**, 610-615.
Ropers, H. H. and Hamel, B. C. (2005). X-linked mental retardation. *Nat. Rev. Genet.* **6**, 46-57.
Rosenthal, N. and Brown, S. (2007). The mouse ascending: perspectives for human-disease models. *Nat. Cell Biol.* **9**, 993-999.
Rosse, C. and Mejino, J. L., Jr (2003). A reference ontology for biomedical informatics: the foundational model of anatomy. *J. Biomed. Inform.* **36**, 478-500.
Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J. et al. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat. Biotechnol.* **25**, 1251-1255.
Smith, C. L., Goldsmith, C. A. and Eppig, J. T. (2005). The mammalian phenotype ontology as a tool for annotating, analyzing and comparing phenotypic information. *Genome Biol.* **6**, R7.
Sprague, J., Bayraktaroglu, L., Bradford, Y., Conlin, T., Dunn, N., Fashena, D., Frazer, K., Haendel, M., Howe, D. G., Knight, J. et al. (2008). The Zebrafish Information Network: the zebrafish model organism database provides expanded support for genotypes and phenotypes. *Nucleic Acids Res.* **36**, D768-D772.
Stenson, P. D., Ball, E., Howells, K., Phillips, A., Mort, M. and Cooper, D. N. (2008). Human gene mutation database: towards a comprehensive central mutation database. *J. Med. Genet.* **45**, 124-126.
Sundberg, J. P. (1994). The hairless (hr) and rhino (hrrh) mutations, chromosome 14. In *Handbook of Mouse Mutations with Skin and Hair Abnormalities: Animal Models and Biomedical Tools* (ed. J. P. Sundberg), pp. 291-312. Boca Raton: CRC Press.
Sundberg, J. P., Dunstan, R. W. and Compton, J. G. (1989). *Hairless Mouse, HRS/J hr/hr.* Heidelberg: Springer-Verlag.
Sundberg, J. P., Price, V. H. and King, L. E. (1999). The "hairless" gene in mouse and man. *Arch. Dermatol.* **135**, 718-720.
Sundberg, J. P., Silva, K. A., Li, R., King, L. E. and Cox, G. A. (2004). Adult onset alopecia areata is a complex polygenic trait in the C3H/HeJ mouse model. *J. Invest. Dermatol.* **123**, 294-297.
Thorisson, G. A., Lancaster, O., Free, R. C., Hastings, R. K., Sarmah, P., Dash, D., Brahmachari, S. K. and Brookes, A. J. (2009). HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.* **37**, D797-D802.
Tiffin, N., Adie, E., Turner, F., Brunner, H. G., van Driel, M. A., Oti, M., Lopez-Bigas, N., Ouzounis, C., Perez-

Iratxeta, C., Andrade-Navarro, M. A. et al. (2006). Computational disease gene identification: a concert of methods prioritizes type 2 diabetes and obesity candidate genes. *Nucleic Acids Res.* **34**, 3067-3081.

van Driel, M. A. and Brunner, H. G. (2006). Bioinformatics methods for identifying candidate disease genes. *Hum. Gen.* **2**, 429-432.

van Driel, M. A., Bruggeman, J., Vriend, G., Brunner, H. G. and Leunissen, J. A. (2006). A text-mining analysis of the human phenome. *Eur. J. Hum. Genet.* **14**, 535-542.

von Linne, C. and Schroeder, J. (1763). *Genera Morborum.* Uppsala: Steinert.

Washington, N. L., Haendel, M. A., Mungall, C. J., Ashburner, M., Westerfield, M. and Lewis, S. E. (2009). Linking human diseases to animal models using ontology-based phenotype annotation *PLoS Biol.* **7**, e1000247.

World Health Organisation (2008). *International Statistical Classification of Diseases and Health Related Problems (The) ICD-10*. Geneva: WHO.

Yuan, R., Tsaih, S. W., Petkova, S. B., de Evsikova, C. M., Xing, S., Marion, M. A., Bogue, M. A., Mills, K. D., Peters, L. L., Bult, C. J. et al. (2009). Aging in inbred strains of mice: study design and interim report on median lifespans and circulating IGF1 levels. *Aging Cell* **8**, 277-287.